Shuyao Wang

STAT 222

Individual write-up

**Introduction**

Alzheimer's disease is a prevalent form of dementia that usually commences with minor memory loss, eventually leading to impaired conversational skills and the inability to react to one's surroundings. The initial damage transpires in regions of the brain related to memory, including the hippocampus and entorhinal cortex, followed by affecting parts of the cerebral cortex that regulate language, reasoning, and social behavior. In due course, other areas of the brain are impacted, resulting in cognitive deterioration. The root cause of this disease is often the abnormal aggregation of proteins, such as tau and amyloid, around brain cells, which leads to inflammation, tissue damage, and cognitive decline.

Currently, three assessments are available to diagnose Alzheimer's disease. Firstly, clinical assessment involves scrutinizing symptoms, medical history, and medication history, and interviewing a close friend or family member. Second, neuropsychological assessment is a thorough evaluation of cognitive abilities, including memory and thinking. Lastly, neuroimaging assessment, such as MRI, utilizes magnetic fields and radio waves to produce detailed images of the brain, indicating regions where shrinkage has occurred. PET scans also detect abnormal protein buildup and reveal normal and abnormal chemical activity in the brain.

Various biomarkers signify Alzheimer's disease diagnosis, including enlarged ventricle size, hippocampal atrophy, and sulci shrinkage. These biomarkers help in accurately diagnosing the disease, which is critical for selecting appropriate treatment options and developing preventive measures.

According to recent data, in 2020, a substantial 5.8 million American individuals were afflicted by Alzheimer's disease, and this number is expected to double every 5 years beyond age 65. The cost of managing the disease is expected to soar to between $379 and more than $500 billion annually by 2040. Therefore, it is crucial for us to establish early diagnosis protocols to reduce the expenses of medical and long-term care for families and the U.S. government. Early detection also plays a significant role in preventing the advancement of the disease to a critical stage, and it allows people to take appropriate measures early on, increasing their chances of benefiting from treatment.

The objective of our project is to utilize time series prediction to aid in the early diagnosis of Alzheimer's disease. To accomplish this, we will be using a dataset downloaded from the Alzheimer's Disease Neuroimaging Initiative (ADNI), which is a research program focused on the progression of Alzheimer's disease. The dataset includes a wide range of information sources, including medical imaging techniques such as magnetic resonance imaging (MRI) and positron emission tomography (PET), cognitive tests, genetics, cerebrospinal fluid (CSF) biomarkers, and blood biomarkers. This comprehensive dataset will enable us to effectively predict the progression of Alzheimer's disease and contribute to the advancement of early diagnosis.

We have identified several specific steps to achieve this goal. Our first step is to predict the likelihood of the subject being infected with Alzheimer's disease within a year. If the subjects are infected, we aim to predict which stage of Alzheimer's they are currently in and how long it will take for them to transition into another stage. According to Elaheh et al., MCI patients will convert to AD over a 3-year period (2015), and we are going to predict whether it is true or not in the project.

**Data**

The present study utilized data sourced from the Alzheimer's Disease Neuroimaging Initiative (ADNI), a research program that aims to bring together researchers and institutions to study the progression of Alzheimer's disease. This initiative aims to collect, validate, and utilize data from a diverse range of sources to understand the changes that occur in the brain during the development and progression of Alzheimer's disease. The data collection process involved a variety of methods such as medical imaging techniques like magnetic resonance imaging (MRI) and positron emission tomography (PET), cognitive tests, genetics, cerebrospinal fluid (CSF) biomarkers, and blood biomarkers. Researchers involved in ADNI collected and validated this data from participants with Alzheimer's disease, mild cognitive impairment, and elderly controls. The collected data is then used to define the progression of Alzheimer's disease and develop better ways to predict, diagnose, and treat the disease.

Our group got the data by directly applying from the official website of ADNI. The data we got consists of the header which is a block of metadata information that precedes the image data in the file and the image files. The header provides essential information on the dimensions, data type, orientation, and other properties of the image data. Our primary focus was on the baseline data and 3T (n = 199) data due to their respective collection of initial MRI data and improved image quality.

To process the data, we initially used Google Colab but later transferred it to the group SCF, which applied through the Department of Statistics at UC Berkeley, given the raw file size exceeded 12GB. For the neuroimaging data, we began our preprocessing steps by first unzipping all the NIFTI files into memory and reading them into the ANTs objective which includes the mask to read the image. We used several packages, including zipfile, matplotlib, squarify, seaborn, numpy, pandas, nibabal, nilearn, ANTS, and pytorch, to read and process the data.

**EDA**

We first examined the metadata and discovered that 199 subjects had imbalanced data between CN and AD (Fig1) and we find that actually, men subjects have more Mild Cognitive Impairment (Fig2). Therefore, we decided to perform data augmentation at a later stage using the mix-up method with NC and AD to generate the MCI class, as the illness is a gradual process. Our MCI baseline will be based on the weighted mean of matrices of NC and AD. We may also propose the cutout method as another data augmentation method in the future, whereby a small square is removed from the image (smaller than ROI), allowing the algorithm to recognize the remaining picture and classify it accurately.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

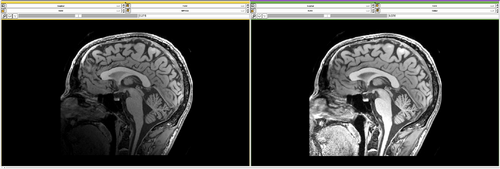
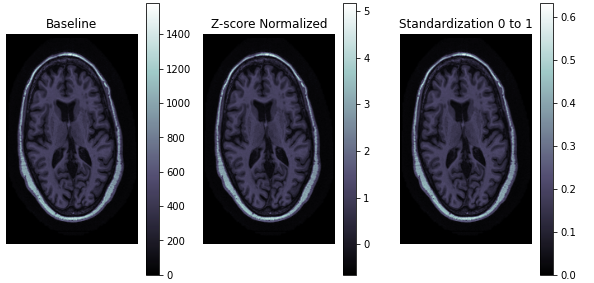
Chart, treemap chart

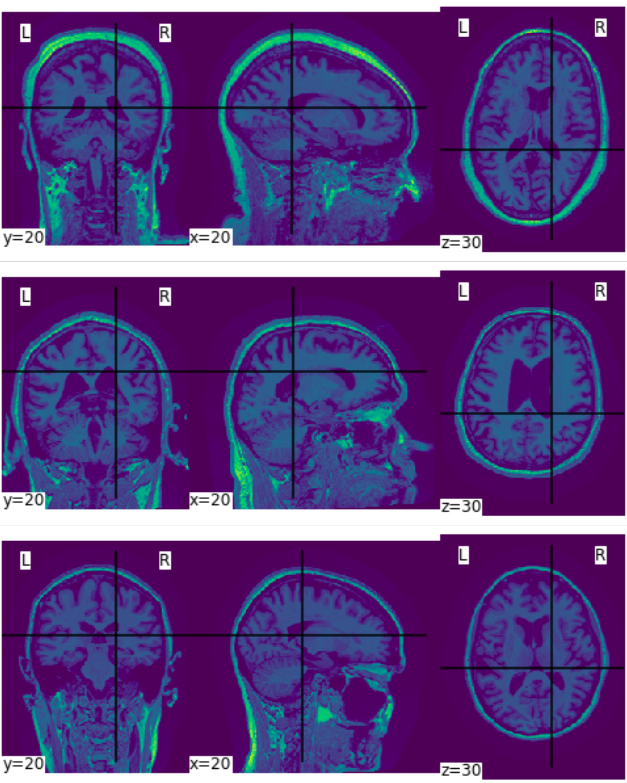
Description automatically generatedFig1.Cognitive impairment classification Fig2. Cognitive impairment classification by sex

Moreover, 31.78% of the data is not unique (Fig3). This means some of our data will be of the exact same person. A person could have 5 MRI scans of them taken, or 2. Having the same photo taken of a person but 5 days apart, could be useless and make cause bias in the model. We might have to remove them later or find a new dataset.

Fig3.Unique number of MRI image

We found that the neuroimaging data for 199 subjects had three image dimensions that needed to do image resampling to fit the DNN model. Besides, we also performed N4 correction (Fig4) that corrects for intensity non-uniformity in the image and intensity normalization (Fig5) that rescales the intensity values of the images to a common range [0,1] after storing them in an array. We used the BEaST library to produce the mask and stripped the image from the raw image which has a 1 mm/voxel resolution. We then manually edited out the skull, scalp, fat, muscle, and other areas. Here are some graphs that illustrate our ideas.

The images we got finally look like this (Fig6). 

Fig4. Intensity Correction Fig5. Intensity Normalization

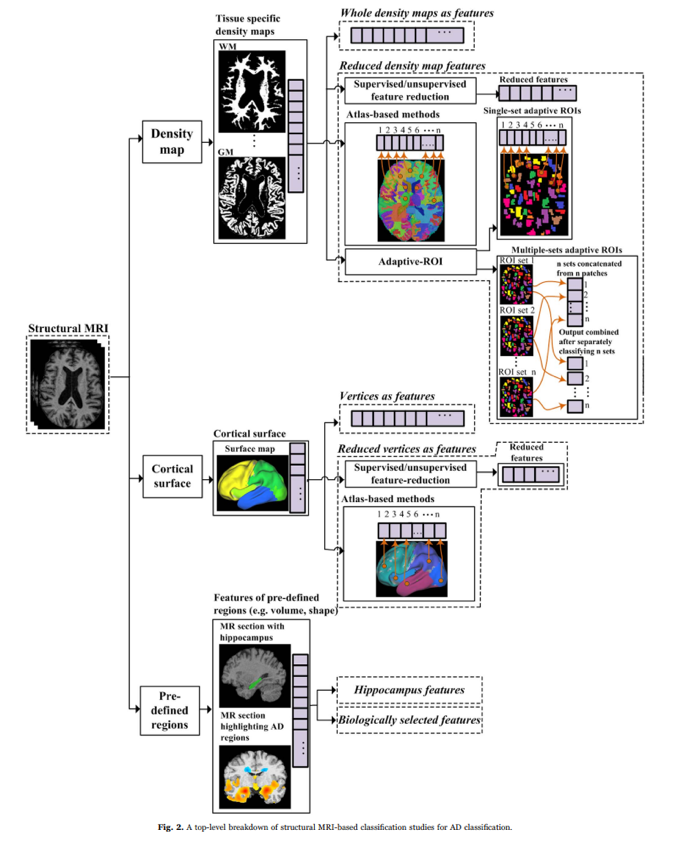


Fig6. Different slices of brain images Fig7. a top-level breakdown of structural MRI-based classification studies for AD classification.

**Model**

In this study, we employed a pre-trained ResNet50 model (Fig7) for the classification of brain images. The ResNet50 model is a well-established model that has been trained on ImageNet. We selected this model because it serves as an effective nonlinear dimensional reduction technique by projecting the data onto a lower dimensional space. Our approach involved selecting the brain slice image with the highest variance among all the scan slices of each individual in order to transform the 3D data into 2D. We then fine-tuned the ResNet50 model and used it to obtain a low-dimensional embedding of the 2D image. Subsequently, we trained a multilayer perceptron (MLP) using the ResNet50 model's output, which resulted in a binary classification output. This approach proved to be effective in accurately classifying brain images. To achieve this, we used several packages, including ResNet50, PyTorch, NumPy, and Scikit-learn.

**Plan for the rest**

The current study aims to undertake Task 1 and Task 2 sequentially in the months of March and April, respectively. Additionally, various interesting techniques will be employed, and essential preprocessing steps such as motion correction, affine transformation, registration, smoothing, and skull stripping will be completed at a later stage. For Task 1, classical traditional methods and a few DL-based methods will be fitted, followed by inference and a particular emphasis on transfer learning. As for Task 2, the analysis will commence with the "Pooled" estimator, and subsequently, temporal CNN/transformer will be explored, along with other potential methods to address the problem.

**Reference**

Elaheh Moradi, Antonietta Pepe, Christian Gaser, Heikki Huttunen, & Jussi Tohka. (2015). Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. NeuroImage, 104, 398-412. doi:10.1016/j.neuroimage.2014.10.002